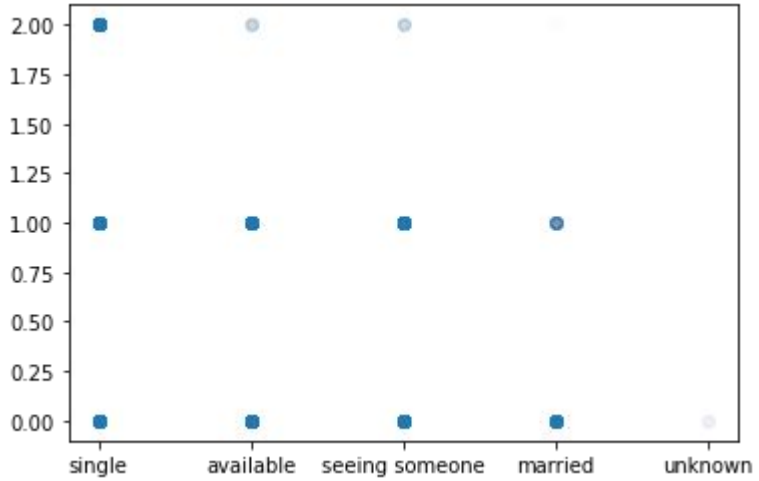
Codecademy ML Capstone

Exploration of the dataset

I plotted relationship status vs drug use (drug use enumerated from 0-2)

The interesting (albeit a bit expected) here is that indeed most drug users are single and almost none of them are married



Exploration of the dataset

I plotted mean income vs jobs (filtering out income == -1 and job == ‘rather not say’

Retired and unemployed people have the highest income for some reason? Otherwise, execusives and financial sector has most money, which makes sense. But then I expected artists to earn less than most, which isn’t the case here.

education / academia	49326.92308
construction / craftsmanship	51363.63636
sales / marketing / biz dev	60683.76068
other	61693.98907
student	62702.7027
political / government	68928.57143
clerical / administrative	74642.85714
entertainment / media	85800
medicine / health	89186.04651
transportation	97000
hospitality / travel	109811.3208
computer / hardware / software	111666.6667
science / tech / engineering	113916.6667
military	130833.3333
law / legal services	132903.2258
artistic / musical / writer	137177.9141
executive / management	139215.6863
banking / financial / real estate	139696.9697
unemployed	153636.3636
retired	342500

Coming up with questions

Looking at the column names, I was curious to see if there is any real relationship between body types of people and their diet, as well as their drinking/smoking habits. I turned this into a classification problem.

As for regression, there wasn’t much data to regress for me, however I was curious to see if I could infer age from income and job data, as well as again drinking/smoking habits. The reasoning behind was that the job field may determine the person’s age group, as well as their drinking/smoking habits.

Body type classification

For this problem, I created a new column for a person being a meat eater or not. And called it `is_veg` (is vegetarian/vegan).

I also created `drinks_code`, `smokes_code`, `drugs_code` for enumerated options for drinking/smoking/doing drugs.

As for body types, I divided it into 3 groups, skinny ('thin', 'skinny', 'used up'), average ('average', 'full figured', 'fit', 'athletic', 'jacked') and overweight (the rest).

Body type classification

Results:

I used DecisionTree and SVC classifiers to try to classify body types from diet, drinking, smoking and drug habits.

SVC: time:0.17s accuracy: 0.68

DecisionTree: time:0.0014s accuracy: 0.67

Since around 68% of body types were average, the classifiers didn’t do any better than always guessing ‘average’ as an answer. However, SVC was a lot more slower than DecisionTree.

Age regression

For this problem, I created a one hot encoding of jobs, since they aren’t linearly related. (I used pd.get_dummies)

I also slightly modified income i.e if the income is -1, I made it 0.

I used the `drinks_code`, `smokes_code`, `drugs_code` columns from the classification problem.

Age regression

Results:

I used SVM and KNN regressors for this problem.

SVM: time:61.55s r2 score: 0.101

KNN: time:129.42s (running the loop of neighbors 1 to 20) r2 score: 0.112

The r2 scores are too low to indicate there is a relation between age and income, job type, and drinking/smoking/drug habits. SVM is faster than KNN if number of neighbors being used isn’t known for KNN.

Conclusion

After working with the data, I couldn’t manage to find anything substantial that wasn’t already known before. My starting point was to print the correlation table of the dataset, and seeing no substantial correlation there, I was doubtful that I will find any good result using regression/classification, and I indeed couldn’t. I’m curious to see what the other students of this skill path found out.